**Predicting Hospital Length of Stay**

**Hyewon Choi, Weixin Peng, Kai Jones, Kathryn Addabbo**

**Abstract:**

Increasing healthcare costs in an unstable political and social landscape in the United States requires better understanding of how costs arise. Patient length of stay (LoS) in a hospital is one factor that directly contributes to increased healthcare costs. Models predicting length of stay can help better understand the factors that influence LoS, such as changes in management, administration and policy. To complete such an analysis, a dataset containing information on 3612 patients was used, which was collected by the Good Health Corporation. Upon reviewing this data, a multiple linear regression model was produced with LoS as the outcome variable with eight predictors, validated by the bootstrap method. Summary statistics for this model include an adjusted r2 value of 0.1385. As expected, predictive elements were all determinants of the patient’s health. However due to the low r2 value, it is suggested that a different modeling technique be used to explore the variables responsible for the variation in LoS. Upon understanding the complex relationship between the different covariates, management and administration can be altered to reduce cost.

**Introduction**

High healthcare costs are of primary concern in the United States (Ravi B. Parikh, 2017). Cost is consistently rising with questionable improvements in quality of care in comparison to other nations (Joseph L. Dieleman, 2016). Due to this and systemic concerns, there has been widespread debate on healthcare policy and how to optimize management and healthcare administration (Barack H. Obama, 2017) (Clinton, 2016) (Henry J. Aaron, 2017). To do this, a better understanding of factors that influence overall costs is necessary. For example, patient length of stay (LoS) in a hospital, is directly associated with cost (Paul A Taheri, 2000) (Fine MJ, 2000). There have been various attempts at predicting LoS using statistical modelling (A. Azari, 2013) (Gordon H. Robinson, 1966). Good Health Corporation (GHC) has collected data on 3612 hospital patients including LoS and other variables, requesting that a predictive model with LoS as the outcome be created based on input predictors.

**Methods**

**Data Source**

GHC collected a total of 3682 visit records on 3612 patients who were admitted to the hospital and over the age of 17. The visit had to have occurred within 24 hours of hospital admission. Data collected for each visit included: length of stay in the hospital in days, the modified early warning score (MEWS), the Charlson Comorbidity Index rank, if the patient had an ICU visit during hospitalization, the number of ER visits in the previous 6 months, the patient’s insurance type, patient demographics, and patient vital signs.

**Data Processing and Cleaning**

The GHC dataset was processed and cleaned to ensure data accuracy and therefore model validity. If a patient had more than one visit, only the first visit was included for model building. Due to skew and nonnormality of the LoS outcome variable, LoS was natural logarithm transformed. For vital sign predictor variables, outliers were identified using the standard z-score method, where values outside of the middle 99.9% of the distribution, or 3.291 standard deviations away from the mean, were replaced with the mean for the predictor. For O2 saturation all values over 100% were removed. This processing removed unrealistic values such as temperatures over 50degrees Celsius. For model building, we only used relevant predictors including: 30 day readmit rate, ER visits in past 6 months, Charson Index rank, MEWS, ICU visit during hospitalization, demographics (age, race, marital status), insurance type and vital signs (respiration rate, O2 saturation, BMI, heartrate, temperature, diastolic blood pressure and systolic blood pressure). We decided to omit data on patient religion due to lack of relevance.

**Model Selection**

Using our selected predictors, we utilized both stepwise regression and criterion based automatic procedures to select the best multiple linear regression model. The final model with the highest adjusted R2 value, lowest CP value and as few predictors as possible to ensure usability was selected.

**Model Diagnostics**

Once the optimal model was selected, assumptions were evaluated. The residuals v. fitted value plot detects error heteroscedasticity and the quantile-quantile plot normality of residuals. The scale-location plot evaluates residual spread and the residuals v. leverage plot helps identify influential cases. Model outliers in LoS were screened for using studentized residuals. To evaluate all possibilities, outliers were removed and regression was rerun. This resulted in an adjusted r2 value and it was decided to keep the initial model. Leverage values in predictors were screened for and none were significant enough to remove. Finally, multicollinearity was evaluated using VIF values with the resulting findings as not significant.

**Model Validation**

The final model was validated using the bootstrap method with 1000 repeats and calculated bias estimates for the model coefficients.

**Results:**

**Data Summary**

The mean length of stay was 5.461 days with a standard deviation of 5.92. Loss of data occurred from cleaning in addition to base missing data for certain predictors (viewed in table 1 as smaller n). Summary statistics for all continuous parameters used in model selection are included in table 1 and proportions for categorical variables are included in table 2.

**Final Model**

The final model meets all assumptions and is given by criterion based automatic model selection. Predictors include: readmit rate within past 30 days, ER visits, Charson Index rank, age, respiration rate, heartrate, temperature and systolic blood pressure. Model coefficients are included in table 3. The adjusted r2 for the model is X the CP score is X and the AIC value is X. All model diagnostic graphs are included in figure 1. Bootstap validation produced biased values for all model coefficients which are included in table 3, which were low and non-significant.

**Discussion**

A predictive model was built for hospital length of stay using 8 predictors and included a log transformation for the outcome variable. Therefore each coefficient in table 3 must be controlled for with all other variables. Most predictors were related to vital signs or current health status. This suggests that length of stay in the hospital is associated with how healthy the patient is or if they have recently visited the hospital (ER visits and admission to a hospital within the past 30 days). The model suggests that factors such as insurance type, race or marital status have negligible impact on hospital stay. It is worth noting that the model does not fit the data well with an adjusted r2 of 0.1385, which is inflated based on replacement of outliers in numeric variables with the mean of said variable. Though bootstrap validation suggested that the final model is valid, it may not be the best possible model. Utilizing other methods such as different regressions (quadratic, exponential, etc.) or advanced machine learning techniques (artificial neural nets, deep learning, etc.) may produce a better predictive model.

# References

A. Azari, V. P. (2013). Predicting Hospital Length of Stay (PHLOS): A Multi-tiered Data Mining Approach. *IEEE Explore*.

Barack H. Obama, J. (2017). Repealing the ACA without a Replacement — The Risks to American Health Care. *The New England Journal of Medicine*, 376:297-299.

Clinton, H. (2016). My Vision for Universal, Quality, Affordable Health Care. *The New England Journal of Medicine*, 375:e36.

Fine MJ, P. H. (2000). Relation between length of hospital stay and costs of care for patients with community-acquired pneumonia. *The American Journal of Medicine*, 378-85.

Gordon H. Robinson, L. E. (1966). Prediction of Hospital Length of Stay. *Health Services Research*, 287–300.

Henry J. Aaron, P. (2017). Which Road to Universal Coverage? *The New England Journal of Medicine*, 377:2207-2209.

Joseph L. Dieleman, R. B. (2016). US Spending on Personal Health Care and Public Health, 1996-2013. *JAMA*, 316(24):2627-2646.

Paul A Taheri, D. A. (2000). Length of stay has minimal impact on the cost of hospital admission. *Journal of the American College of Surgeons*, 123-130.

Ravi B. Parikh, M. M. (2017). Getting Real about Health Care Costs — A Broader Approach to Cost Stewardship in Medical Education. *The New England Journal of Medicine*, 376:913-915.

**Tables and Figures:**

**Table 1.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **n** | **Mean** | **Sd** | **Minimum** | **Maximum** | **Median** |
| Length of Stay (Days) | 3532 | 5.404 | 5.804 | 0.04167 | 87.96 | 3.792 |
| Age (Years | 3532 | 65.68 | 18.69 | 18 | 105 | 68 |
| ER Visits in past 6 months | 3532 | 1.743 | 1.577 | 0 | 4 | 1 |
| BMI | 2849 | 27.94 | 5.869 | 5.1 | 51.9 | 28.35 |
| Systolic BP | 3527 | 130.5 | 16.55 | 88.78 | 184.9 | 129.3 |
| Diastolic BP | 3531 | 72.22 | 8.901 | 43.65 | 104.2 | 71.86 |
| O2 Saturation | 3529 | 97.38 | 1.596 | 80 | 100 | 97.54 |
| Temperature (C) | 3530 | 36.74 | 0.4135 | 33.78 | 39.65 | 36.73 |
| Heart Rate | 3527 | 79.68 | 11.9 | 37.58 | 122.4 | 79.06 |
| Respiration Rate | 3529 | 3529 | 1.598 | 12 | 26.79 | 17.75 |

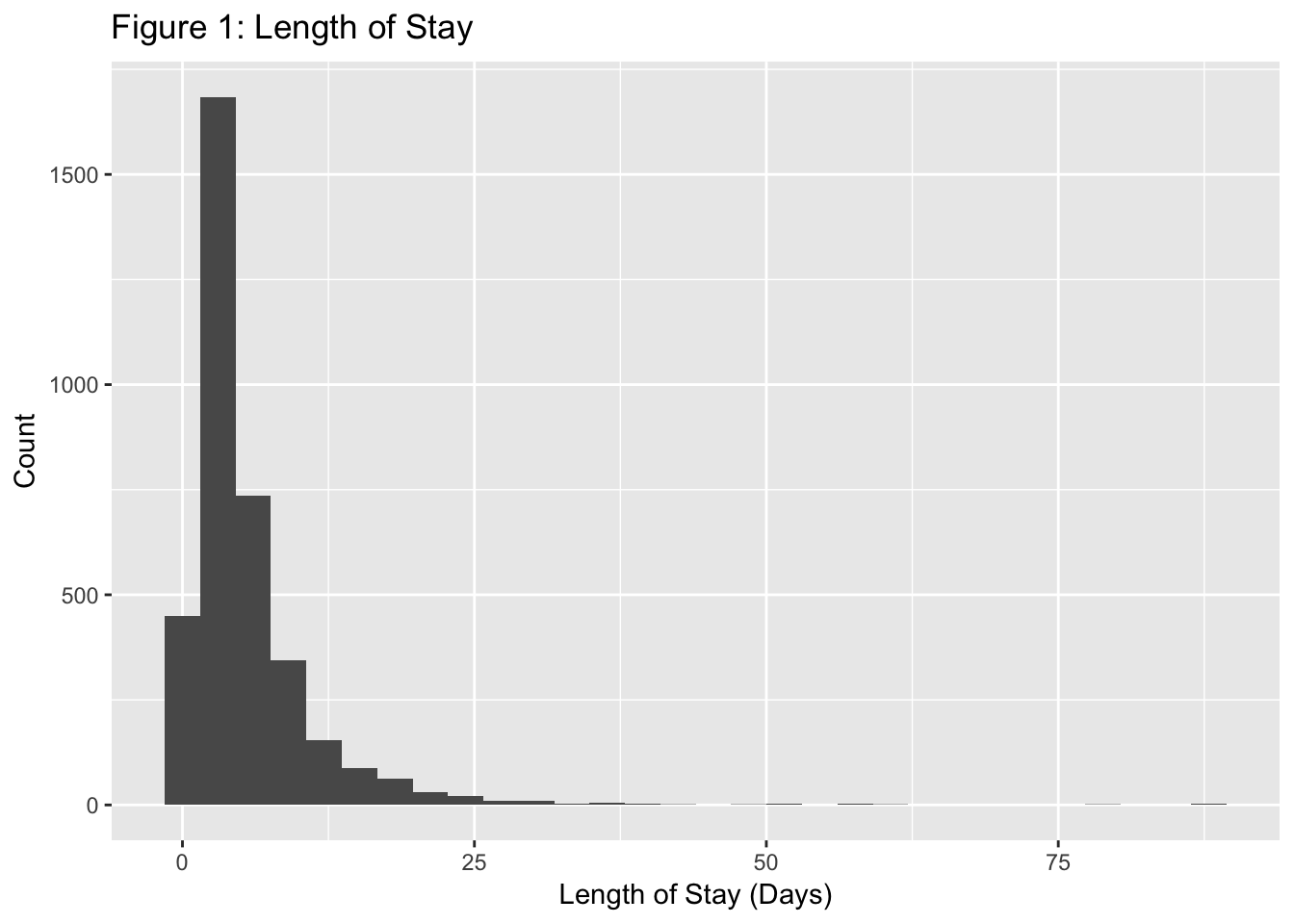
**Table 2.**

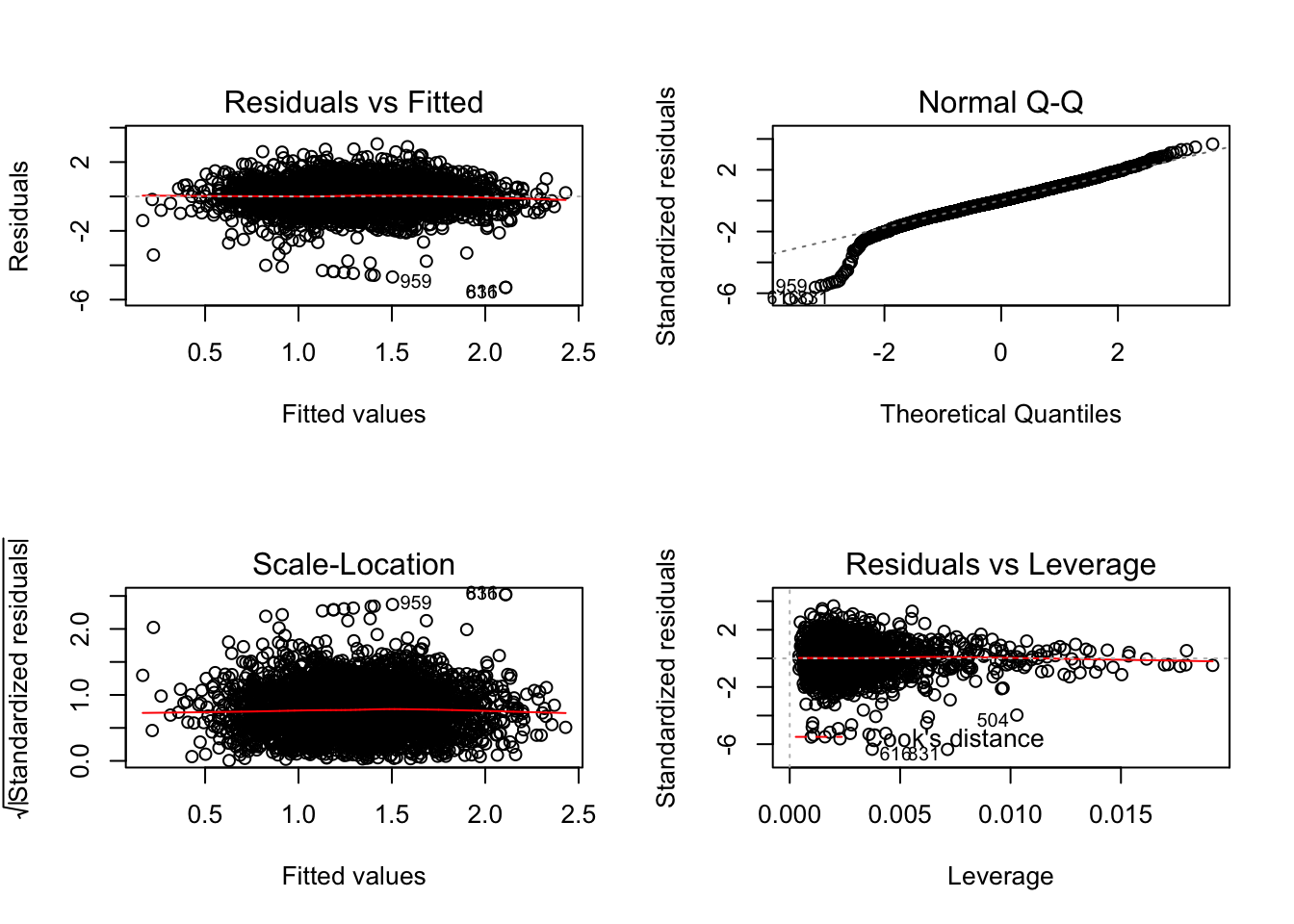
|  |  |
| --- | --- |
| **Variable** | **N (%)** |
| **Gender** | |
| Male | 1660 (46.0) |
| Female | 1952 (54.0) |
| **Race** | |
| White | 2057 (56.9) |
| Black | 772 (21.4) |
| Asian | 249 (6.9)\_ |
| Native American (Alaskan) | 22 (0.6) |
| Native American (Hawaiian/Pacific Islander) | 4 (0.1) |
| Other | 508 (14.1) |
| **Insurance Status** | |
| Medicare | 1425 (39.5) |
| Medicaid | 166 (4.6) |
| Private | 1987 (55.0) |
| Missing Data | 34 (0.9) |
| **Marital Status** | |
| Single | 951 (26.3) |
| Married | 1607 (44.5) |
| Widowed | 690 (19.1) |
| Divorced | 235 (6.5) |
| Separated | 51 (1.4) |
| Civil Union | 1 (0.0) |
| Missing Data | 77 (2.1) |
| **30 Day Readmit Rate** | 517 (14.3) |

**Table 3**

|  |  |  |  |
| --- | --- | --- | --- |
| Predictor | Coefficient Value | P-Value | Bootstrap Bias |
| Intercept | -6.6121133 | 3.98e-07 | -4.421041e-02 |
| Readmit in Past 30 Days | 0.1807683 | 2.80e-05 | 4.331755e-04 |
| ER Visits in past 6 months | 0.0669272 | 4.55e-12 | 5.771404e-04 |
| Charson Index rank | 0.0421502 | 5.93e-07 | -3.036915e-04 |
| Age (Years) | 0.0109857 | < 2e-16 | 1.558764e-05 |
| Respiration Rate | 0.0660580 | 8.66e-13 | 1.109412e-03 |
| Heartrate | 0.0076618 | 1.66e-09 | 1.178587e-05 |
| Internal Temperature | 0.1616098 | 4.61e-06 | 4.391052e-04 |
| Systolic BP | -0.0058352 | 1.49e-10 | 3.853057e-05 |

**Figure 1. Length of Stay**

****

**Figure 2. Residuals**